

Washington University School of Medicine

Digital Commons@Becker

2020-Current year OA Pubs

Open Access Publications

7-1-2021

Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: A systematic literature review

Sayantana Kumar

Inez Oh

Suzanne Schindler

Albert M Lai

Philip R O Payne

See next page for additional authors

Follow this and additional works at: https://digitalcommons.wustl.edu/oa_4





Part of the [Medicine and Health Sciences Commons](#)

Authors

Sayantana Kumar, Inez Oh, Suzanne Schindler, Albert M Lai, Philip R O Payne, and Aditi Gupta

Review

Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: a systematic literature review

Sayantana Kumar¹, Inez Oh¹, Suzanne Schindler², Albert M. Lai ¹, Philip R.O. Payne ¹, and Aditi Gupta¹

¹Institute for Informatics, Washington University School of Medicine, St. Louis, Missouri, USA and ²Department of Neurology, Washington University School of Medicine, St. Louis, Missouri, USA

Corresponding Author: Aditi Gupta, PhD, Institute for Informatics, Washington University School of Medicine, Campus Box 8102, 660 S Euclid Ave, St. Louis, MO 63110, USA; agupta24@wustl.edu

Received 14 April 2021; Revised 21 June 2021; Editorial Decision 22 June 2021; Accepted 30 June 2021

ABSTRACT

Objective: Alzheimer disease (AD) is the most common cause of dementia, a syndrome characterized by cognitive impairment severe enough to interfere with activities of daily life. We aimed to conduct a systematic literature review (SLR) of studies that applied machine learning (ML) methods to clinical data derived from electronic health records in order to model risk for progression of AD dementia.

Materials and Methods: We searched for articles published between January 1, 2010, and May 31, 2020, in PubMed, Scopus, ScienceDirect, IEEE Explore Digital Library, Association for Computing Machinery Digital Library, and arXiv. We used predefined criteria to select relevant articles and summarized them according to key components of ML analysis such as data characteristics, computational algorithms, and research focus.

Results: There has been a considerable rise over the past 5 years in the number of research papers using ML-based analysis for AD dementia modeling. We reviewed 64 relevant articles in our SLR. The results suggest that majority of existing research has focused on predicting progression of AD dementia using publicly available datasets containing both neuroimaging and clinical data (neurobehavioral status exam scores, patient demographics, neuroimaging data, and laboratory test values).

Discussion: Identifying individuals at risk for progression of AD dementia could potentially help to personalize disease management to plan future care. Clinical data consisting of both structured data tables and clinical notes can be effectively used in ML-based approaches to model risk for AD dementia progression. Data sharing and reproducibility of results can enhance the impact, adaptation, and generalizability of this research.

Key words: Alzheimer disease, dementia, electronic health records, clinical data, machine learning

INTRODUCTION

Alzheimer disease (AD) is the most common cause of dementia, which is a syndrome characterized by impairment of memory and/or thinking severe enough to interfere with activities of daily life.^{1,2} AD dementia affects millions of people worldwide and is currently the sixth leading cause of death in the United States.^{3,4} AD-related brain

pathology, which includes the accumulation and deposition of amyloid- β peptide and tau protein, begins almost 10–20 years before the onset of dementia symptoms.⁵ Therefore, many individuals with early AD brain pathology are cognitively normal but at higher risk for developing dementia in the future.⁶ AD dementia is progressive and incurable, and, at advanced stages, patients suffer potentially

LAY SUMMARY

Alzheimer disease (AD) is the most common cause of dementia, which is a syndrome of impaired memory and/or thinking that interferes with activities of daily life. Many studies of AD dementia utilize information from expensive and invasive procedures, such as brain imaging or spinal taps, to estimate risk for developing AD dementia or rapid cognitive decline. However, widely available electronic health records (EHRs) systems contain a wealth of healthcare data describing a patient's medical history and clinical presentation (eg, demographics, vital signs, medications, laboratory data, current medical conditions) that could be leveraged as a low-cost, noninvasive alternative to study the progression of AD dementia. In recent years, machine learning (ML) has become a useful tool in identifying hidden patterns within large-scale healthcare data, such as the aforementioned EHR-derived data types, leading to increased efficiency and improved healthcare. We aim to perform a systematic literature review of studies applying ML to clinical and EHR data to identify factors that predict risk for progression of AD dementia. We summarize the reviewed articles according to key components of ML analysis such as data characteristics, computational algorithms, and research focus. Finally, we identify gaps in the literature and potential opportunities for future research.

fatal complications such as dehydration, malnutrition, or infection.⁷ Identifying individuals with early AD brain pathological changes could lead to therapeutic interventions to delay the disease progression over time and could be helpful for tailoring disease management and planning future care.

Clinical data are defined as “information ranging from determinants of health and measures of health and health status to documentation of care delivery . . . captured for a variety of purposes and stored in numerous databases across the healthcare system.”⁸ Nonimaging clinical data extracted from EHRs are some of the most accessible and widely used clinical datasets. They are an integral part of contemporary healthcare delivery, enabling quick access to accurate, up-to-date, and complete patient information, and assisting in accurate diagnosis and coordinated, efficient care.⁹ EHR data collected from individuals at risk for AD dementia can include laboratory test results, vital signs, medications, and other treatments administered, as well as comorbidities.^{9,10} In some cases, patients may also undergo specific testing for markers linked to AD brain pathology using expensive and/or invasive procedures such as neuroimaging scans (magnetic resonance imaging [MRI] and position emission tomography [PET]) and cerebrospinal fluid (CSF) collection for biomarker testing.^{11–15} The results of these tests may also be present in the EHR. Research has shown that such longitudinal clinical EHR data (ie, data collected from multiple time points) can be utilized for monitoring the time-course of AD dementia progression.¹⁶

The widespread use and availability of medical devices over the years have provided an overwhelming volume of clinical EHR data, which could potentially augment the traditional tools of dementia experts.¹⁷ The unmet needs for dementia expertise, coupled with the relevant massive datasets, have encouraged researchers to examine the utility of artificial intelligence (AI), which is gaining high visibility in the realm of healthcare innovation.¹⁸ Machine learning (ML), a branch of AI, can model the relationship between the input quantities and clinical outcomes, discover hidden patterns within large-scale data, and make inferences or decisions that help in more accurate clinical decision-making.¹⁹ However, computational hypotheses generated by ML models still need to be validated by subject matter experts in order to ensure adequate precision for clinical decision-making purposes.²⁰ In our review, we include studies using ML for the purpose of predictive modeling (such as decision trees, support vector machines [SVM], k-means clustering) and exclude studies using statistical methods for cohort summarization and hypothesis testing (such as odds ratio, Chi-square distribution, Kruskal–Wallis

test, Kappa-Cohen test). For studies using linear and logistic regression only those studies were included which utilized these methods for predictive modeling or classification analysis.

The current literature generally neglects secondary use of nonimaging clinical data, including routinely collected EHR-derived data, as a rich, low-cost, and noninvasive source of information for identifying potential risk factors for AD dementia, and instead focuses on the use of costlier and/or invasive imaging and diagnostic testing data for ML-based analysis. However, we believe that further study of EHR-derived data could lead to more efficient, cost-effective, timely, and personalized disease management for individuals with AD. With this motivation and a goal of identifying the knowledge gaps and potential opportunities for the use of EHR-derived data in conjunction with ML frameworks, our goal was to conduct a systematic literature review (SLR) on the state-of-the-art of ML as applied EHR-derived data for the purposes of modeling and understanding AD dementia progression.

METHODS

As noted above, the motivation behind conducting an SLR is to summarize existing findings related to a chosen research topic to identify gaps in literature and thus create a ground for future research work. As stated by Martí-Juan et al,¹⁹ “Performing an SLR comprises the following steps: (1) identify the need for performing the SLR; (2) formulate research questions; (3) execute a comprehensive search and selection of primary studies; (4) assess the quality and extract data from the studies; (5) interpret the results; and (6) report the SLR.” In this SLR, our main research question is: *How are machine learning algorithms being applied by researchers for studying progression of AD dementia using clinical EHR data?* This main question can be subdivided further into the following 3 research questions:

1. What type of ML methods have been used for detecting the onset of AD dementia and for predicting the trajectory of the disease progression?
2. What EHR-derived data types and risk factors (eg, physiological, genetics, demographics) have been used as features for predictive modeling?
3. What are the research foci of the reviewed articles that use ML methods on EHR-derived data for modeling and predicting the progression of AD dementia?

Search strategy

The aim of this SLR is to review works that meet the following criteria: (1) focus on modeling and predicting the onset or progression of AD dementia; (2) use ML techniques; and (3) use clinical markers of patients diagnosed with AD dementia. Similar to ref, [19](#) we created 3 keyword groups as follows, each relevant to different aspects of the scope of the review.

1. Keywords related to disease: AD, Alzheimer's, Alzheimer, Alzheimer's disease, dementia, Alzheimer's Disease, and Related Dementia.
2. Keywords related to ML methodology: ML, machine learning, AI, artificial intelligence, pattern recognition, computer-aided-diagnosis, CAD, classification, prediction, supervised learning, unsupervised learning, predictive modeling.
3. Keywords related to data and features: electronic health records, EHR, clinical data, clinical assessments, patient health data.

For each of the 3 keyword groups, we selected the words as per standard terminologies used for manuscript notation in the targeted literature databases.^{[19,21](#)} The disease group consisted of different words related to AD and dementia. We observed that the terms "Mild Cognitive Impairment" and "MCI" often relate to brain diseases other than AD, so we excluded those to avoid false positives. For the ML methodology group, we focused on all possible variants related to ML/AI as well as general terms like prediction, classification, etc. The third group related to data and features consisted of keywords related to clinical EHR data. Since the focus of the review was on the use of clinical EHR-derived data with/without imaging features, keywords related to imaging like "neuroimaging," "MRI," "PET," "CT," etc., were not part of any inclusion or exclusion criteria.

For our SLR, we searched the following bibliographic databases: (1) PubMed; (2) Association for Computing Machinery Digital Library; (3) IEEE Explore Digital Library; (4) ScienceDirect; (5) Scopus; and (6) arXiv/BioarXiv. Works originally identified from arXiv/BioarXiv were subsequently verified to have been accepted in a peer-reviewed journal or conference. Using each search engine, we searched the titles, abstracts, and keyword sections of articles published in journals or conference proceedings between January 1, 2010, and May 31, 2020. In order to limit our search to the scope of the review, our search string in each of the online databases was a triplet with 1 keyword from each of the 3 groups. All possible string combinations were created by taking 1 term from each of the 3 keyword groups joined by an "AND." The set of above formed triplets were then used as queries for the search.

Exclusion criteria

The entire procedure of article searching and inclusion/exclusion criteria was performed according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.^{[22](#)} The inclusion/exclusion criteria were reviewed by a board-certified neurologist and dementia specialist. All articles were selected and screened for eligibility by a doctoral student. In the first phase of screening, all duplicate articles collected from the different source libraries were removed. The next phase of screening discarded all papers that were clearly not relevant to the review, including studies where the abstract and the title did not contain any of the keywords related to "Alzheimer disease" or "Machine Learning." Following the approach adopted by the authors in ref,^{[21](#)} the articles that passed the screening phase were assessed for eligibility by reviewing

the full texts of the remaining articles to exclude studies which met one or more of the exclusion criteria ([Table 1](#)).

Study risk of bias assessment

To mitigate the risk of bias during the search process and inclusion/exclusion of the articles, a series of checks were implemented during the article selection process. All articles were selected and screened for eligibility by a doctoral student. Two additional authors validated the final set of papers and review analysis. The final selected articles and inclusion/exclusion criteria were also reviewed by a board-certified neurologist and dementia specialist for relevance to our main research question.

Summary statistics

Replicating the method followed in ref, [21](#) we calculated the following summary statistics from the final set of included articles: (1) source and publication year of article; (2) research focus of the article; (3) modality and accessibility of dataset; (4) size of cohort and type of features/risk factors; and (5) type of ML model for predictive modeling.

RESULTS

[Figure 1](#) shows the PRISMA flowchart, which depicts the selection process by which we arrived at the final set of included studies. We identified a total of 1331 studies from the different bibliographic databases. Since many papers were included in multiple databases, the first exclusion step removed 405 duplicate articles. From the remaining 926 articles, we removed 345 studies that were out of the scope of our review. This includes articles where the title or abstract did not contain any of the keywords related to "Alzheimer disease" or "Machine Learning." We reviewed 581 full-text articles for eligibility based on the exclusion criteria described in [Table 1](#). After filtering on one or more of the exclusion criteria in [Table 1](#), 64 articles remained for review.

[Figure 2](#) shows the distribution of the reviewed articles by their publication year. The value shown for the year 2020 only includes data until May, when we performed the search. As shown by the plot, the count of published papers relevant to our scope has increased over the past 5 years, demonstrating that interest in using ML for analyzing AD with clinical data is on an upward trajectory.

Data characteristics

To understand the nature of the data used by researchers in their articles, we documented the accessibility of the dataset, the number of included human subjects, and the incorporated clinical features. For each of the articles, we checked if the authors provided directions on how to access the dataset used in their experiments. We observed 2 main categories of datasets from our analysis: (1) deidentified datasets that are publicly available for download and (2) restricted datasets from sources like institutional clinical datasets that are not available for public use.

For the first category, we found that the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset was the most widely used, with 64% (41/64) of articles using ADNI for longitudinal AD data. ADNI enrolls participants between the ages of 55 and 90 at sites in the United States and Canada. After obtaining informed consent, participants undergo a series of neuropsychological and clinical assessments, genetic testing, and imaging (MRI and PET) at multiple

Table 1. Exclusion criteria for research articles

	Exclusion criteria	Reasons for exclusion
1.	Only neuroimaging features were considered for predictive modeling	Scope of the review was inclusion of clinical EHR-derived data with/or without imaging features
2.	ML methods were not used for clinical predictive modeling related to AD/dementia	We excluded articles which performed cohort summarization and hypothesis testing using statistical methods like logistic regression odds ratio, Chi-square distribution, Kruskal–Wallis test, etc.
3.	Focus on a disease other than AD/dementia	AD/dementia is not the focus of the main analyses
4.	AD/dementia is used only as an example of a neurodegenerative disease	AD/dementia is not the focus of the main analyses
5.	Not peer-reviewed conference proceedings, journal, or pre-prints	Outside the scope of our review
6.	Multiple publications from the same research group with similar final outcomes. In such cases, only the most recent studies were considered	Considered to be duplicate articles
7.	Review articles	Review articles did not focus on a specific research goal

Abbreviations: AD: Alzheimer disease; EHR: electronic health record; ML: machine learning.

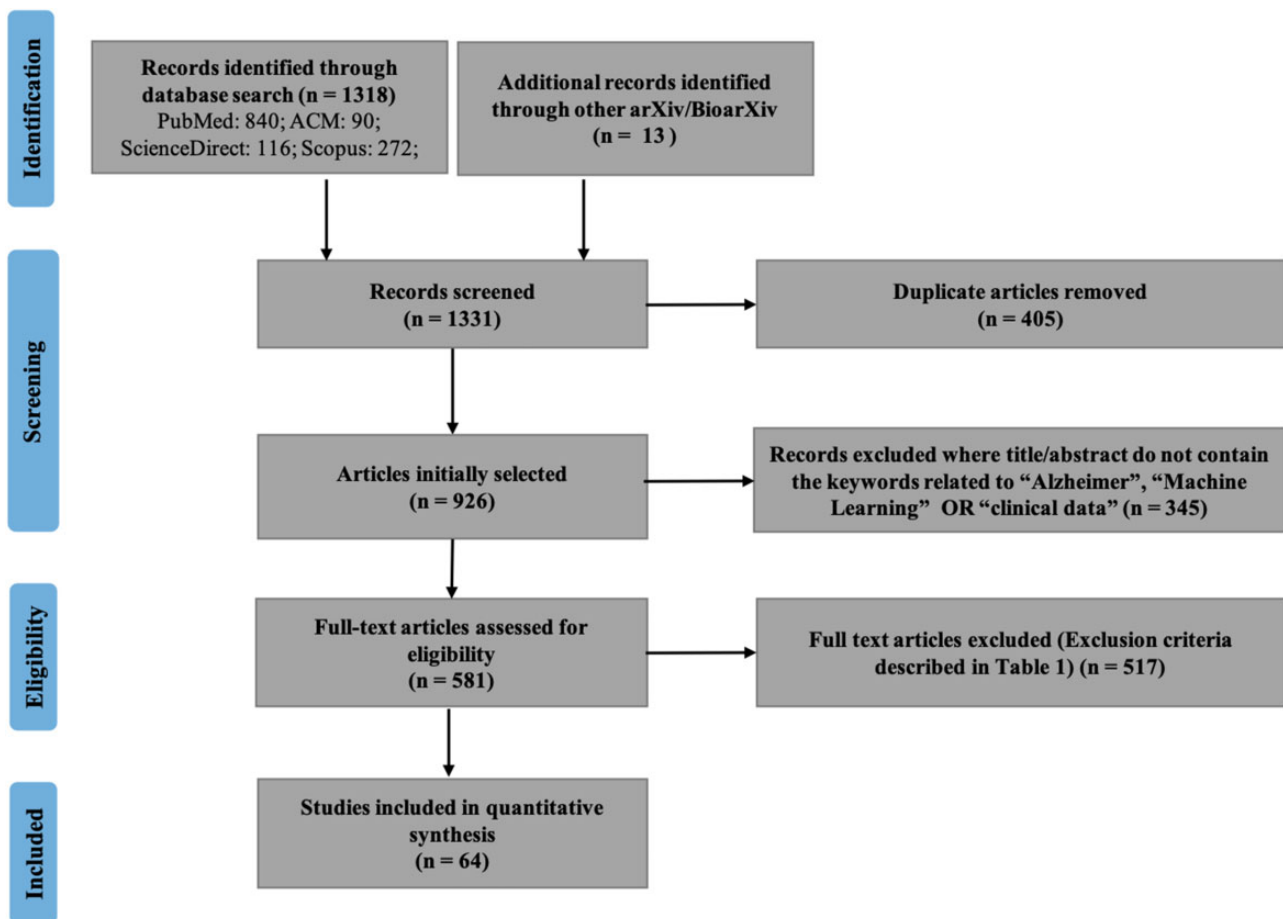


Figure 1. PRISMA flow diagram. Abbreviation: ACM: Association for Computing Machinery.

timepoints over subsequent years.²³ Other data sources included the National Alzheimer's Coordinating Center;^{24–26} Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing;²⁷ Framingham Heart Study;^{28,29} and Coalition Against Major Databases.³⁰ These datasets are all publicly available for download with some requiring a license from their respective websites.

For the second category, 16 out of 64 papers used their own customized clinical datasets. We deemed such datasets “restricted data” when there were no references or external links through which the data could be accessed. Examples of restricted datasets analyzed in these papers included a dataset subsampled from the National Health Insurance Service—national sample cohort of 1 million

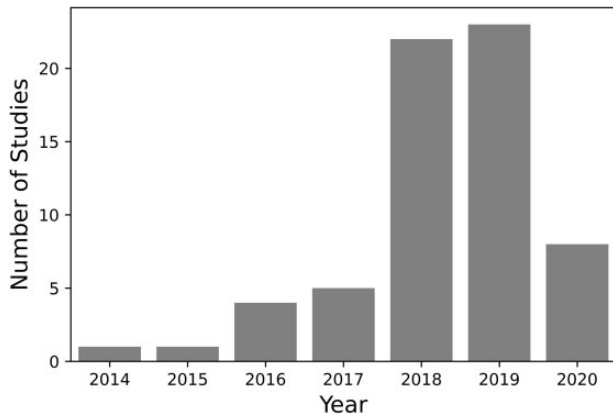


Figure 2. Studies published per year which use machine learning on clinical data for prognostic estimates of Alzheimer's Disease. For 2020, studies dated till May 31st were considered for the review.

people representative of the South Korean population within the Korean National Health Insurance Service database³¹ and a dataset collected from people who received a screening test in the dementia center in Gangbuk-Gu, Seoul, from 2008 to 2013.³²

The performance of an ML framework depends significantly on the size of the training cohort. From each of the reviewed articles, we determined the number of patients whose health data were used for analysis of AD dementia. Based upon the observed cohort sizes in all the included studies, we divided the cohort sizes into 4 categories: (1) 0–1000 patients; (2) 1001–10 000; (3) 10 001–100 000; and (4) >100 000. Figure 3 shows the number of studies corresponding to each cohort size category. Three articles^{33–35} did not report the cohort sizes; so, we excluded them from our analysis of Figure 3. Most of the studies had cohort sizes of 0–1000 patients ($n = 34$), followed by 1001–10 000 patients ($n = 15$), 10 001–100 000 patients ($n = 5$), and finally >100 000 patients ($n = 7$).

AD dementia features and biomarkers

Data of patients analyzed for risk of AD dementia consist of clinical variables like laboratory results, vital signs, neurobehavioral status exam scores, demographic information, and comorbidities, along with neuroimaging scans and CSF biomarkers. ML models try to learn the relationships amongst a set of clinical variables and determine if and how these variables contribute to the model predicting the development of AD dementia. Although the scope of this review

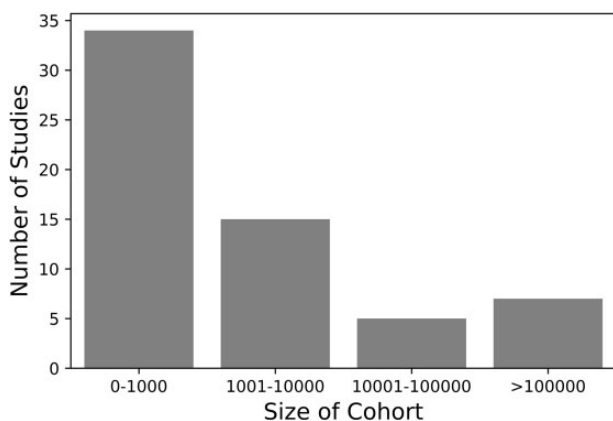


Figure 3. Size of cohort used in the reviewed studies.

focused on articles using nonimaging clinical data, some of these studies used multimodal datasets with nonimaging clinical and imaging features. To determine the prevalence of the use of nonimaging clinical features as potential AD risk factors, we classified the included studies into the following 2 categories: (1) Clinical only—only nonimaging clinical variables^{36,37} and (2) Clinical + Imaging—imaging and nonimaging clinical variables were integrated to form the complete set of features.^{38–41} We grouped the features into the following categories: neuroimaging features,^{35,42,43} cognitive assessments,^{44–47} genetic factors,^{48–51} laboratory test values,^{52–54} patient demographics,^{55–57} and clinical notes.

Table 2 summarizes the feature categories and measures/factors used while applying the ML framework along with the count of articles using that particular variable. In our cohort of studies that used data from neuroimaging techniques as features for predictive modeling, MRI was the most widely used.^{35,42,43} As shown in Table 2, cognitive assessments (48 studies) and demographics (47 studies) were the 2 most common features used by researchers for analysis of AD dementia. Only 4 articles considered clinical notes, primarily patient medical history and diagnosis details documented by clinicians.^{58–61} Thirty studies (47%) were categorized as Clinical only and 34 (53%) as Clinical + Imaging. Figure 4 shows the relationship between the nature of data access restrictions (publicly available or restricted) and the category of AD dementia features (Clinical only or Clinical + Imaging). As illustrated by the figure, 94% (32/34) of the Clinical + Imaging data were extracted from publicly available datasets. For the Clinical only studies, 57% (17/30) of the Clinical data originated from datasets which are not publicly accessible.

Application of ML methods

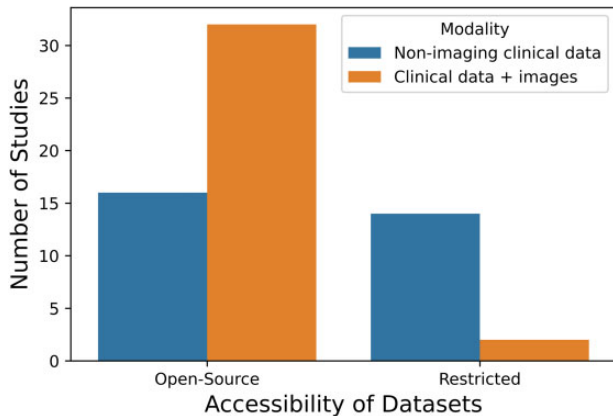
ML tools can model complex relationships between different clinical variables that are often beyond human capabilities. The output of trained ML models when applied on previously unseen healthcare clinical data yield inferences that can augment clinical decision-making. With the advancement of computational resources, researchers have progressed from simple ML algorithms like regression to complex deep learning models. We examined the different categories of ML techniques used in the reviewed articles based on the model type and the type of learning algorithm. We grouped the ML methods based on model type into the following categories: regression,^{52,62,63} SVM,^{64–66} decision tree,^{67–69} Bayesian networks,^{70–72} neural networks,^{33,73,74} and natural language processing (NLP). The neural networks category includes both classifiers such as multilayer perceptron and deep learning models such as convolutional neural networks and autoencoders.^{75–77} Based on the type of learning algorithm, ML models can either be supervised, unsupervised, or semi-supervised. In a supervised learning model, the algorithm is trained on a dataset annotated with gold standard labels. Unsupervised learning models, on the other hand extract features and patterns from unlabeled data and cluster the data points into distinct classes. Semi-supervised learning is a hybrid of the above 2 methods and combines a small amount of labeled data with a large volume of unlabeled data during training. Regression, SVM, decision trees, Bayesian networks, and neural networks all fall under the supervised category. k-Means algorithm is an example of unsupervised learning.⁶⁰

Table 3 summarizes the types of ML models based on model type along with their different variants. The most widely used techniques were decision trees (50%), neural networks (44%), regression (34%), SVM (34%), and Bayesian networks (20%). NLP was

Table 2. Features related to AD dementia identified by articles

Feature categories	Measures/factors	Number of articles (%)
Neuroimaging	MRI (structural, functional, unspecified), PET (FDG, amyloid)	35 (54%)
Cognitive assessments	MMSE, ADAS-Cog, others (CDR, FAQ)	48 (75%)
Genetic	APOE ϵ 4, family history	24 (38%)
Laboratory	CSF, vitals, medications, medical history, other laboratory tests	32 (50%)
Demographics	Age, gender, education, race	47 (72%)
Clinical notes	Discharge summary	4 (6%)

Abbreviations: AD: Alzheimer disease; ADAS-Cog: Alzheimer's Disease Assessment Scale-cognitive subscale; APOE ϵ 4: apolipoprotein epsilon 4 allele; CDR: clinical dementia rating; CSF: cerebrospinal fluid; FAQ: Functional Activities Questionnaire; FDG, ●●●; MMSE: Mini-Mental State Exam; MRI: magnetic resonance imaging; PET: position emission tomography.

**Figure 4.** Relationship between the modality and accessibility of the datasets used in the included studies.

used only in studies which included patient clinical notes as one of the features; these comprised 6% of all studies.^{58–61}

Research foci of the reviewed articles

Identifying the research foci or the end goals of the reviewed articles can provide insights into how applying ML techniques on clinical EHR data can lead to effective clinical decision-making. More than half of the reviewed articles (55%, 35/64) investigated the progression of AD dementia to determine if an individual had stable or progressive AD. They aimed to predict the development of AD dementia in individuals who were initially cognitively normal or had only mild cognitive impairment.^{34,78,79} Eleven/sixty-four (17%) studies used data comprised of longitudinal trajectories of clinical variables from patients showing mild symptoms of dementia to train predictive models for personalized forecasting of disease progression.^{80–82} Eighteen/sixty-four (28%) studies in our review aimed to satisfy both the above objectives. For example,^{83,84} presented a computational ML-based framework for modeling symptom trajectories using cognitive assessment scores at multiple time points and subsequently predicting those trajectory classes using multimodal data comprising both clinical and imaging variables.

Reproducibility

Reproducibility, a fundamental requirement of the scientific process, is related to the idea that a scientific experiment should be able to be reproduced to validate its results.⁸⁵ The reproducibility of a scientific study is often assessed by the extent to which it follows the FAIR (Findable, Accessible, Interoperable and Reusable) princi-

ples.⁸⁶ For our review, we analyzed if the authors have provided adequate details about the dataset used and the implementation to check if the FAIR principles were followed. Only 7 (11%) of the articles reported their implementation code^{27,37} and 48 (75%) studies used publicly available datasets.^{36,38,41}

DISCUSSION

AD is the most common cause of dementia, which is a syndrome of impaired memory and/or thinking that interferes with activities of daily life. We performed an SLR of studies applying ML to clinical data to identify factors that predict risk for progression of AD dementia. There has been an exponential increase in such paper applying ML for AD over the past 3 years. We reviewed the selected articles according to key components of ML analysis such as data characteristics, computational algorithms, and research focus; in the process we identified gaps in the existing literature and potential opportunities for future research. Our review suggests that most of the articles focus on predicting the progression of the disease based on standardized publicly available multimodal datasets which include both neuroimaging and some nonimaging clinical data. Most commonly used nonimaging clinical features for predictive modeling include neurobehavioral status exam scores, patient demographics, neuroimaging data, and laboratory test values.

Clinical databases which are collected for specific research purposes (eg, data in clinical registries) or cleaned and curated to enhance data reusability (eg, MIMIC database⁸⁷) are often relatively well-structured, standardized, and clean, even though they may still have a few missing values and outliers. Hence, many researchers focus on utilizing these relatively clean datasets for their experiments and methodological innovations. However, as we noted previously, clinical data from local sources like institutional EHRs, which are primarily used to track patient care but can also be used secondarily for clinical research and automated disease surveillance, have great potential for use in modeling AD dementia progression.⁸⁸ Data from such raw EHR data sources often have data quality issues and require significant effort for data preprocessing and feature engineering. However, they are a rich source of historical clinical data containing patient-level elements which can be effectively leveraged using ML-based computational techniques for longitudinal analyses of their preclinical phase to identify prognostic clinical phenotypes, thus representing an opportunity to employ precision medicine paradigms in disease states where the current evidence-base precludes such an approach.

The basic criteria for selecting articles for our review was inclusion of clinical data excluding imaging with/without other features and/or data types. We observed that a significant portion of articles employed neuroimaging features from structural and functional MRI

Table 3: Specific computational and machine learning methods utilized

Computational methods	Specific models	Number of articles (%)
Regression	Linear regression	22 (34%)
	Logistic regression	
	Lasso regression	
	Ridge regression	
	Support vector regression	
SVM	SVM with linear kernel	22 (34%)
	SVM with RBF kernel	
	SVM with polynomial kernel	
	Support vector regression	
Decision trees	Decision trees	32 (50%)
	Random forest	
	Adaboost	
	GBM	
Bayesian networks	Naïve Bayes model	13 (20%)
	Bayesian belief networks	
	GMM	
Neural networks	Multilayer perceptron	28 (44%)
	CNN-based models	
	RNN-based models	
	Autoencoder	
	RBM	
NLP	Graph neural networks	4 (6%)
	Text mining	
Others	KNN	7 (11%)
	k-Means	

Abbreviations: CNN: convolutional neural network; GBM: gradient boosting models; GMM: Gaussian mixture model; KNN: K-nearest neighbor; NLP: natural language processing; RBF: radial basis function; RBM: restricted Boltzmann machines; RNN: recurrent neural network; SVM: support vector machines.

as well as fluorodeoxyglucose (FDG) and amyloid positron emission tomography (PET) for predictive modeling; this indicates that most researchers use clinical features as part of multidimensional datasets containing both clinical and imaging features. As evident from the relationship between the modality and accessibility of the datasets, multimodal features are mostly derived from the category of publicly accessible, standardized, and well-curated datasets.

Identifying individuals with early AD brain pathological changes could enable therapeutic interventions to delay the disease progression over time and can be helpful for tailoring disease management and planning future care. Multiple failed drug trials for AD dementia show that in the later stages of the disease course, when the patient already has significant neuronal degeneration, treatment is unlikely to be helpful.⁸⁹ Hence, many drug trials are now enrolling patients with either preclinical AD (cognitively normal individuals with AD-related brain pathology) or very early AD dementia.^{90,91} Most studies reported that their proposed methodology can identify individuals at risk for progression to AD dementia approximately 24–48 months before the diagnosis of AD dementia.

Nearly all of the reviewed articles used supervised learning in the proposed models. Unsupervised or semi-supervised learning can also be an effective tool for handling multidimensional longitudinal patient data where clinical outcomes are not known *a priori*. Unsupervised clustering algorithms can be helpful for identifying novel subphenotypes with distinct disease trajectories and the associations between them.^{89,92}

EHR-derived data for patients who are screened for risk of AD dementia not only include structured data in the form of labs, medications, and procedures, but also clinical notes, which are textual descriptions of physician–patient encounters and records of their follow-up visits.⁹³ We observed from the summary statistics that in-

formation from clinical notes are not often included for developing the predictive modeling pipelines. However, these notes often consist of additional clinical information that are usually unavailable in the structured data sources, offering a rich source of information for clinical decision-making. Most of the notes are free-text narratives lacking a standardized structure and they cannot be processed by conventional ML algorithms like SVM, decision trees, regression, etc. NLP is a field of computational techniques that offers a viable solution for effectively processing clinical notes. In recent years, deep learning-based NLP models like recurrent neural networks and long short-term memory networks have been shown to outperform the conventional word-embedding-based NLP techniques for extracting relevant information from clinical notes.⁹⁴

Data sharing and reproducibility of results can also enhance the impact, adaptation, and generalizability of research. Ideally, measures such as use of standardized publicly available EHR-derived datasets and specification of implementation details can help ensure reproducibility of the published methods and results. However, siloed data between different academic and corporate institutions and inconsistent data formats often make data sharing difficult and therefore remains an open area of research and innovation.⁹⁵ A solution to this problem is developing a culture of data sharing among different institutions, potentially utilizing common data models like the OHDSI (Observational Health Data Sciences and Informatics) data sharing initiative. OHDSI produces tools like the OMOP (Observational Medical Outcomes Partnership) Common Data Model, which transforms data within disparate observational databases into a common representation (terminologies, vocabularies, and coding schemes) so that they can be analyzed using standardized analytics tools.⁹⁶

Despite these limitations, there have been significant advancements in the application of ML on EHR-derived data for predicting

AD dementia progression over the last 2–3 years. Advancements in technology and computational tools provide an opportunity for researchers to develop deep learning-based computational hypotheses that can inform clinical decision-making. Deep learning and other analytic approaches in ML can define clinical patterns and generate insights beyond human capabilities. This not only reduces the burden on clinicians in making their diagnoses but leads to improved quality, safety, and outcomes of care planning and delivery.¹⁸

Limitations of SLR

Many relevant research works might be published in only conferences and workshops proceedings and not indexed in bibliographic databases. Similarly, some studies from arXiv/BioarXiv were not included since they were not yet peer-reviewed. Thus, identification of only peer-reviewed studies from bibliographic databases might lead to selection bias during the initial inclusion stage. This can potentially impact the results presented as it may not truly represent the growing interest in the domain of using ML models for AD prediction using clinical data.

CONCLUSION

We performed an SLR of studies using ML on clinical EHR data for modeling and prediction of AD dementia progression. We summarized different aspects of the articles including data source and modality, features, methods, and research focus of the studies. The summarized results suggest that most state-of-the-art research on AD has focused on predicting the progression of the disease based on standardized publicly available multimodal datasets which include both neuroimaging and some nonimaging clinical data. The nonimaging clinical data used most commonly for predictive modeling include neurobehavioral status exam scores, patient demographics, neuroimaging data, and laboratory test values. Almost all the reviewed articles utilized supervised learning with common ML models such as neural networks, decision trees, SVM, and regression. ML and other analytic approaches in AI can generate helpful insights about complex clinical patterns that assists clinicians in their decision-making and leads to improved quality, safety, and outcomes of healthcare planning.

FUNDING

The preparation of this report was supported by the Centene Corporation contract (P19-00559) for the Washington University-Centene ARCH Personalized Medicine Initiative. SS is supported by K23AG053426.

AUTHOR CONTRIBUTIONS

SK and AG conceived and designed the study. SK did the data collection (literature review), analysis, and interpretation, drafted and revised the manuscript, and prepared the graphical illustrations. AG and IO participated in the literature review interpretation, drafted, and revised the manuscript, and approved the final version for submission. PROP, AML, and SS reviewed and revised the manuscript, and approved the final version for submission.

SUPPLEMENTARY MATERIAL

[Supplementary material](#) is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The data underlying this article are available in the article and in its online [supplementary material](#).

REFERENCES

- McKhann GM, Knopman DS, Chertkow H, *et al.* The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 2011; 7 (3): 263–9.
- Ballard C, Gauthier S, Corbett A, *et al.* Alzheimer's disease. *Lancet* 2011; 377 (9770): 1019–31.
- Ferri CP, Prince M, Brayne C, *et al.* Global prevalence of dementia: a Delphi consensus study. *Lancet* 2005; 366 (9503): 2112–7.
- Prince M, Bryce R, Albanese E, *et al.* The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimers Dement* 2013; 9 (1): 63–75.e2.
- Villemagne VL, Burnham S, Bourgeat P, *et al.*; Australian Imaging Biomarkers and Lifestyle (AIBL) Research Group. Amyloid beta deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. *Lancet Neurol* 2013; 12 (4): 357–67.
- Musiek ES, Schindler SE. Alzheimer disease: current concepts & future directions. *Mo Med* 2013; 110 (5): 395–400.
- Förstl H, Kurz A. Clinical features of Alzheimer's disease. *Eur Arch Psychiatry Clin Neurosci* 1999; 249 (6): 288–90.
- McGinnis JM, Olsen LM, Goolsby WA. *Clinical Data as the Basic Staple of Health Learning: Creating and Protecting a Public Good: Workshop Summary*. Washington, DC: National Academies Press; 2011.
- Vaughn VM, Linder JA. *Thoughtless Design of the Electronic Health Record Drives Overuse, but Purposeful Design Can Nudge Improved Patient Care*. *BMJ Qual Saf* 2018; 27 (8): 583–586.
- Knopman DS, DeKosky ST, Cummings JL, *et al.* Practice parameter: diagnosis of dementia (an evidence-based review). Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology* 2001; 56 (9): 1143–53.
- Bilgel M, Jedynak B, Wong DF, Resnick SM, Prince JL. Temporal trajectory and progression score estimation from voxelwise longitudinal imaging measures: application to amyloid imaging. In: *International Conference on Information Processing in Medical Imaging*. Skye, Scotland: Springer; 2015.
- Chi C-L, Zeng W, Oh W, *et al.* Personalized long-term prediction of cognitive function: using sequential assessments to improve model performance. *J Biomed Inform* 2017; 76: 78–86.
- Li K, Luo S. Functional joint model for longitudinal and time-to-event data: an application to Alzheimer's disease. *Stat Med* 2017; 36 (22): 3560–72.
- Shaw LM, Arias J, Blennow K, *et al.* Appropriate use criteria for lumbar puncture and cerebrospinal fluid testing in the diagnosis of Alzheimer's disease. *Alzheimers Dement* 2018; 14 (11): 1505–21.
- Johnson KA, Minoshima S, Bohnen NI, *et al.* Appropriate use criteria for amyloid PET: a report of the Amyloid Imaging Task Force, the Society of Nuclear Medicine and Molecular Imaging, and the Alzheimer's Association. *Alzheimers Dement* 2013; 9 (1): e1–16.
- Mills KL, Tamnes CK. Methods and considerations for longitudinal structural brain imaging analysis across development. *Dev Cogn Neurosci* 2014; 9: 172–90.
- Dallora AL, Eivazzadeh S, Mendes E, *et al.* Machine learning and microsimulation techniques on the prognosis of dementia: a systematic literature review. *PLoS One* 2017; 12 (6): e0179804.

18. Maddox TM, Rumsfeld JS, Payne PR. Questions for artificial intelligence in health care. *JAMA* 2019; 321 (1): 31–2.
19. Martí-Juan G, Sanroma-Guell G, Piella G. A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer's disease. *Comput Methods Programs Biomed* 2020; 189: 105348.
20. Chen P-HC, Liu Y, Peng L. How to develop machine learning models for healthcare. *Nat Mater* 2019; 18 (5): 410–4.
21. Pellegrini E, Ballerini L, Valdes Hernandez MDC, et al. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. *Alzheimers Dement* 2018; 10 (1): 519–35.
22. Moher D, Liberati A, Tetzlaff J, et al.; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009; 6 (7): e1000097.
23. Wu Y, Zhang X, He Y, et al. Predicting Alzheimer's disease based on survival data and longitudinally measured performance on cognitive and functional scales. *Psychiatry Res* 2020; 291: 113201.
24. Khan A, Usman M. Early diagnosis of Alzheimer's disease using informative features of clinical data. In: Proceedings of the International Conference on Machine Vision and Applications. 2018; Singapore.
25. Lin M, Gong P, Yang T, et al. Big data analytical approaches to the NACC dataset: aiding preclinical trial enrichment. *Alzheimer Dis Assoc Disord* 2018; 32 (1): 18–27.
26. Wang T, Qiu RG, Yu M. Predictive modeling of the progression of Alzheimer's disease with recurrent neural networks. *Sci Rep* 2018; 8 (1): 9161.
27. Bhagwat N, Viviano JD, Voineskos AN, et al.; Alzheimer's Disease Neuroimaging Initiative. Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data. *PLoS Comput Biol* 2018; 14 (9): e1006376.
28. Ang TFA, An N, Ding H, et al. Using data science to diagnose and characterize heterogeneity of Alzheimer's disease. *Alzheimers Dement (N Y)* 2019; 5: 264–71.
29. Joshi PS, Heydari M, Kannan S, et al. Temporal association of neuropsychological test performance using unsupervised learning reveals a distinct signature of Alzheimer's disease status. *Alzheimers Dement* 2019; 5 (1): 964–973.
30. Fisher CK, Smith AM, Walsh JR, et al.; Coalition Against Major Diseases. Machine learning for comprehensive forecasting of Alzheimer's Disease progression. *Sci Rep* 2019; 9 (1): 13622.
31. Park JH, Cho HE, Kim JH, et al. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *NPJ Digit Med* 2020; 3 (1): 46–7.
32. So A, Hooshyar D, Park K, et al. Early diagnosis of dementia from clinical data by machine learning techniques. *Appl Sci* 2017; 7 (7): 651.
33. Cao P, Liu X, Liu H, et al. Generalized fused group lasso regularized multi-task feature learning for predicting cognitive outcomes in Alzheimer's disease. *Comput Methods Programs Biomed* 2018; 162: 19–45.
34. Kebets V, Richiardi J, Van Assche M, et al. Predicting pure amnesic mild cognitive impairment conversion to Alzheimer's disease using joint modeling of imaging and clinical data. In: 2015 International Workshop on Pattern Recognition in NeuroImaging. Stanford, CA: IEEE; 2015.
35. Zhang H, Zhu F, Dodge HH, et al.; the Alzheimer's Disease Neuroimaging Initiative. A similarity-based approach to leverage multi-cohort medical data on the diagnosis and prognosis of Alzheimer's disease. *GigaScience* 2018; 7 (7): gij085.
36. Grassi M, Rouleaux N, Caldirola D, et al.; Alzheimer's Disease Neuroimaging Initiative. A novel ensemble-based machine learning algorithm to predict the conversion from mild cognitive impairment to Alzheimer's disease using socio-demographic characteristics, clinical information and neuropsychological measures. *Front Neurol* 2019; 10: 756.
37. Nori VS, Hane CA, Crown WH, et al. Machine learning models to predict onset of dementia: a label learning approach. *Alzheimers Dement (N Y)* 2019; 5: 918–925.
38. Ezzati A, Lipton RB; for the Alzheimer's Disease Neuroimaging Initiative. Machine learning predictive models can improve efficacy of clinical trials for Alzheimer's disease 1, 2. *J Alzheimers Dis* 2020; 74 (1): 55–9.
39. Goyal D, Tjandra D, Migrino RQ, et al.; Alzheimer's Disease Neuroimaging Initiative. Characterizing heterogeneity in the progression of Alzheimer's disease using longitudinal clinical and neuroimaging biomarkers. *Alzheimer Dement* 2018; 10 (1): 629–637.
40. Moore PJ, Lyons TJ, Gallacher J, et al.; Alzheimer's Disease Neuroimaging Initiative. Random forest prediction of Alzheimer's disease using pairwise selection from time series data. *PLoS One* 2019; 14 (2): e0211558.
41. Spasov S, Passamonti L, Duggento A, et al.; Alzheimer's Disease Neuroimaging Initiative. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *NeuroImage* 2019; 189: 276–287.
42. Huang L, Jin Y, Gao Y, et al.; Alzheimer's Disease Neuroimaging Initiative. Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest. *Neurobiol Aging* 2016; 46: 180–191.
43. Yao D, Calhoun VD, Fu Z, et al. An ensemble learning system for a 4-way classification of Alzheimer's disease and mild cognitive impairment. *J Neurosci Methods* 2018; 302: 75–81.
44. Battista P, Salvatore C, Castiglioni I. Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: a machine learning study. *Behav Neurol* 2017; 2017: 1850909.
45. Fulton L, Dolezel D, Harrop J, et al. Classification of Alzheimer's disease with and without imagery using gradient boosted machines and ResNet-50. *Brain Sci* 2019; 9 (9): 212.
46. Jin Y, Su Y, Zhou X-H, et al.; Alzheimer's Disease Neuroimaging Initiative. Heterogeneous multimodal biomarkers analysis for Alzheimer's disease via Bayesian network. *EURASIP J Bioinform Syst Biol* 2016; 2016 (1): 12.
47. Martinez-Murcia FJ, Ortiz A, Gorriç J-M, et al. Studying the manifold structure of Alzheimer's Disease: a deep learning approach using convolutional autoencoders. *IEEE J Biomed Health Inform* 2020; 24 (1): 17–26.
48. Brand L, Nichols K, Wang H, Huang H, Shen L; Alzheimer's Disease Neuroimaging Initiative. Predicting longitudinal outcomes of Alzheimer's disease via a tensor-based joint classification and regression model. In: Pacific Symposium on Biocomputing. Hawaii: World Scientific; 2020.
49. Lee G, Nho K, Kang B, et al.; for Alzheimer's Disease Neuroimaging Initiative. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Sci Rep* 2019; 9 (1): 1–12.
50. Saribudak A, Subick AA, Rutta JA, Uyar MÜ. Gene expression based computation methods for Alzheimer's disease progression using hippocampal volume loss and MMSE scores. In: Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2016; Seattle, WA.
51. Zhu F, Panwar B, Dodge HH, et al. COMPASS: a computational model to predict changes in MMSE scores 24-months after initial assessment of Alzheimer's disease. *Sci Rep* 2016; 6: 34567.
52. Bucholc M, Ding X, Wang H, et al. A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. *Expert Syst Appl* 2019; 130: 157–171.
53. Kim H, Chun H-W, Kim S, et al. Longitudinal study-based dementia prediction for public health. *Int J Environ Res Public Health* 2017; 14 (9): 983.
54. Lee G, Kang B, Nho K, et al. MildInt: deep learning-based multimodal longitudinal data integration framework. *Front Genet* 2019; 10: 617.
55. Geifman N, Kennedy RE, Schneider LS, et al. Data-driven identification of endophenotypes of Alzheimer's disease progression: implications for clinical trials and therapeutic interventions. *Alz Res Therapy* 2018; 10 (1): 1–7.
56. Khanna S, Domingo-Fernández D, Iyappan A, et al. Using multi-scale genetic, neuroimaging and clinical data for predicting Alzheimer's disease and reconstruction of relevant biological mechanisms. *Sci Rep* 2018; 8 (1): 1–13.
57. Tunvirachaisakul C, Supasitthumrong T, Tangwongchai S, et al. Characteristics of mild cognitive impairment using the Thai version of the consortium to establish a registry for Alzheimer's disease tests: a multivariate and machine learning study. *Dement Geriatr Cogn Disord* 2018; 45 (1–2): 38–48.

58. Bin-Hezam R, Ward TE. A machine learning approach towards detecting dementia based on its modifiable risk factors. *Int J Adv Comput Sci Appl* 2019; 10 (8): 148–158.
59. McCoy TH, Han L, Pellegrini AM, et al. Stratifying risk for dementia onset using large-scale electronic health record data: a retrospective cohort study. *Alzheimers Dement* 2020; 16 (3): 531–540.
60. Moreira LB, Namen AA. A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia. *Comput Methods Programs Biomed* 2018; 165: 139–149.
61. Uspenskaya-Cadoz O, Alamuri C, Wang L, et al. Machine learning algorithm helps identify non-diagnosed prodromal Alzheimer's disease patients in the general population. *J Prev Alzheimers Dis* 2019; 6 (3): 185–191.
62. Kang MJ, Kim SY, Na DL, et al. Prediction of cognitive impairment via deep learning trained with multi-center neuropsychological test data. *BMC Med Inform Decis Mak* 2019; 19 (1): 231.
63. Lins AJCC, Muniz MTC, Garcia ANM, et al. Using artificial neural networks to select the parameters for the prognostic of mild cognitive impairment and dementia in elderly individuals. *Comput Methods Programs Biomed* 2017; 152: 93–104.
64. Forouzannezhad P, Abbaspour A, Cabrerizo M, Adjouadi M. Early diagnosis of mild cognitive impairment using random forest feature selection. In: 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS), Cleveland, OH; 2018.
65. Segovia F, Bastin C, Salmon E, Górriz JM, Ramírez J, Phillips C. Automatic differentiation between Alzheimer's Disease and mild cognitive impairment combining PET data and psychological scores. In: 2013 International Workshop on Pattern Recognition in Neuroimaging. Philadelphia, PA: IEEE; 2013: 144–147.
66. Tabarestani S, Aghili M, Shojai M, Freytes C, Adjouadi M. Profile-specific regression model for progression prediction of Alzheimer's disease using longitudinal data. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). Orlando, FL: IEEE; 2018: 1353–1357.
67. Almubark I, Chang L-C, Nguyen T, Turner RS, Jiang X. Early detection of Alzheimer's disease using patient neuropsychological and cognitive data and machine learning techniques. In: 2019 IEEE International Conference on Big Data (Big Data). Los Angeles, CA: IEEE; 2019: 5971–5973.
68. Mahyoub M, Randles M, Baker T, Yang P. Effective use of data science toward early prediction of Alzheimer's disease. In: 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). Exter, UK: IEEE; 2018: 1455–1461.
69. Shahbaz M, Ali S, Guergachi A, Niazi A, Umer A. Classification of Alzheimer's disease using machine learning techniques. In: *DATA*. 2019: 296–303; Prague, Czech Republic.
70. Lahmiri S, Shmuel A. Performance of machine learning methods applied to structural MRI and ADAS cognitive scores in diagnosing Alzheimer's disease. *Biomed Signal Process Control* 2019; 52: 414–419.
71. Satone V, Kaur R, Faghri F, Nalls MA, Singleton AB, Campbell RH. Learning the progression and clinical subtypes of Alzheimer's disease from longitudinal clinical data. *arXiv* 2018. abs/1812.00546.
72. Utsumi Y, Guerrero R, Peterson K, Rueckert D, Picard RW. Meta-weighted gaussian process experts for personalized forecasting of AD cognitive changes. In: Machine learning for healthcare conference. Michigan, Ann Arbor: PMLR; 2019: 181–196.
73. Albright J; Alzheimer's Disease Neuroimaging Initiative. Forecasting the progression of Alzheimer's disease using neural networks and a novel pre-processing algorithm. *Alzheimer Dement* 2019; 5 (1): 483–491.
74. An N, Ding H, Yang J, et al. Deep ensemble learning for Alzheimer's disease classification. *J Biomed Inform* 2020; 105: 103411.
75. Candemir S, Nguyen XV, Prevedello LM, et al.; Alzheimer's Disease Neuroimaging Initiative. Predicting rate of cognitive decline at baseline using a deep neural network with multidata analysis. *J Med Imaging (Bellingham)* 2020; 7 (4): 044501.
76. Shmulev Y, Belyaev M. Predicting conversion of mild cognitive impairments to Alzheimer's disease and exploring impact of neuroimaging. In: *GRAIL/Beyond-MIC@MICCAI*. 2018; Granada, Spain.
77. Zhu W, Razavian N. Variationally regularized graph-based representation learning for electronic health records. In: Proceedings of the Conference on Health, Inference, and Learning. 2021; Virtual event.
78. Forouzannezhad P, Abbaspour A, Li C, Cabrerizo M, Adjouadi M. A deep neural network approach for early diagnosis of mild cognitive impairment using multiple features. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). Orlando, FL: IEEE; 2018: 1341–1346.
79. Lee GG, Huang P-W, Xie Y-R, Pai M-C. Classification of Alzheimer's disease, mild cognitive impairment, and cognitively normal based on neuropsychological data via supervised learning. In: TENCON 2019-2019 IEEE Region 10 Conference (TENCON). Kerala, India: IEEE; 2019: 1808–1812.
80. Nie L, Zhang L, Meng L, et al. Modeling disease progression via multi-source multitask learners: A case study with Alzheimer's disease. *IEEE Trans Neural Netw Learn Syst* 2017; 28 (7): 1508–1519.
81. Pölsterl S, Sarasua I, Gutiérrez-Becker B, Wachinger C. A wide and deep neural network for survival analysis from anatomical shape and tabular clinical data. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Würzburg, Germany: Springer; 2019.
82. Zhu Y, Sabuncu MR. A probabilistic disease progression model for predicting future clinical outcome. *arXiv* 2018. abs/1803.05011.
83. Jarrett D, Yoon J, van der Schaar M. Dynamic prediction in clinical survival analysis using temporal convolutional networks. *IEEE J Biomed Health Inform* 2020; 24 (2): 424–436.
84. Pillai PS, Leong T-Y. Modeling multi-view dependence in Bayesian networks for Alzheimer's disease detection. *Stud Health Technol Inform* 2019; 264: 358–362.
85. Mondelli ML, Peterson AT, Gadelha LM. Exploring reproducibility and FAIR principles in data science using ecological niche modeling as a case study. In: *International Conference on Conceptual Modeling*. Salvador, Brazil: Springer; 2019.
86. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016; 3 (1): 160018.
87. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3 (1): 1–9.
88. Munir Shah S, Khan RA. Secondary use of electronic health record: opportunities and challenges. *arXiv* 2020. arXiv: 2001.09479.
89. Yiannopoulou KG, Anastasiou AI, Zachariou V, et al. Reasons for failed trials of disease-modifying treatments for Alzheimer Disease and their contribution in recent research. *Biomedicines* 2019; 7 (4): 97.
90. Sperling RA, Rentz DM, Johnson KA, et al. The A4 study: stopping AD before symptoms begin? *Sci Transl Med* 2014; 6 (228): 228fs13.
91. Dubois B, Hampel H, Feldman HH, et al.; Proceedings of the Meeting of the International Working Group (IWG) and the American Alzheimer's Association on "The Preclinical State of AD"; July 23, 2015; Washington DC, USA. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. *Alzheimers Dement* 2016; 12 (3): 292–323.
92. Giannoula A, Gutierrez-Sacristán A, Bravo Á, et al. Identifying temporal patterns in patient disease trajectories using dynamic time warping: a population-based study. *Sci Rep* 2018; 8 (1): 1–14.
93. Zhou X, Wang Y, Sohn S, et al. Automatic extraction and assessment of lifestyle exposures for Alzheimer's disease using natural language processing. *Int J Med Inform* 2019; 130: 103943.
94. Wu S, Roberts K, Datta S, et al. Deep learning in clinical natural language processing: a methodical review. *J Am Med Inform Assoc* 2020; 27 (3): 457–470.
95. Lustgarten JL, Zehnder A, Shipman W, et al. Veterinary informatics: forging the future between veterinary medicine, human medicine, and One Health initiatives—a joint paper by the Association of Veterinary Informatics (AVI) and the CTSA One Health Alliance (COHA). *JAMIA Open* 2020; 3 (2): 306–317.
96. Hripscak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015; 216: 574–578.